Available online at www.sciencedirect.com

ScienceDirect

Genomics 91 (2008) 259–266

GENOMICS

www.elsevier.com/locate/ygeno

EnsemPro: An ensemble approach to predicting transcription start sites in human genomic DNA sequences

Hong-Hee Won [a,b], Min-Ji Kim [a], Seonwoo Kim [a], Jong-Won Kim [c,*]

[a] *Samsung Biomedical Research Institute, Sungkyunkwan University School of Medicine, Samsung Medical Center, 50 Ilwon-dong, Kangnam-gu, Seoul 135-710, South Korea*

[b] *Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Guseong-dong, Yuseong-gu, Daejeon 305-701, South Korea*

[c] *Department of Laboratory Medicine and Genetics, Sungkyunkwan University School of Medicine, Samsung Medical Center, 50 Ilwon-dong, Kangnam-gu, Seoul 135-710, South Korea*

Received 11 July 2007; accepted 7 November 2007

Abstract

Although several computational methods have been developed to identify transcription start sites (TSSs)/promoters, the computational prediction still needs improvement. Due to low performance, the promoter prediction programs can provide misleading results in functional genomic studies. To improve the prediction accuracy, we propose the use of an ensemble approach, EnsemPro (Ensemble Promoter), which combines the prediction results of the existing promoter predictors. We schematically compared the prediction performance of the currently available promoter prediction programs in an identical evaluating environment, and the results served as a guide for choosing the combined predictors. We applied three representative ensemble schemes—the majority voting, the weighted voting, and the Bayesian approach—for the TSS prediction of hundreds of human genomic sequences. EnsemPro identified the TSSs more precisely than other combining methods as well as the currently available individual predictor programs. The source code of EnsemPro is available on request from the authors.

Keywords: Bioinformatics; Promoter prediction; Transcription start site; Ensemble approach

Decoding of human genomic sequences has been significantly challenging since the Human Genome Project was first launched. The annotation of enumerating DNA sequences is almost impossible without the support of computational methodology. Identification of the promoter regions of targeted sequences has been a major focus of several investigations [1–3]. Predicting the promoter region or the transcription start site (TSS) is very important as it allows one to study the functional roles of genes. It is very expensive and time-consuming to detect TSSs experimentally or manually, and the use of statistical theories and machine learning methods can make it possible to search automatically for TSSs from the genome sequence via a computational algorithm such as neural network, genetic algorithm, and linear discriminant function. As a result, many programs have been designed and developed in the past several years, including NNPP [4], Proscan [5], TSSG/TSSW [6], PromH [7], FirstEF [8], Dragon promoter finder [9], Eponine [10], and Promoter2.0 [11]. Most of the programs identify the promoter region based on the biological features of TSSs such as the TATA box score or CpG islands. More recently, simple consensus methods using a combination scheme have been proposed [12,13]. Investigators need to choose the best program from all of the available programs for applying it to their own sequences of interest. However, the choice of the best prediction program is difficult, as the programs have not been evaluated within the same experimental environment. Therefore, the programs should be evaluated under the same conditions to be able to choose the appropriate program for a particular application [12,13].

The Condorcet Jury Theorem proved that the judgments of a committee are superior to those of individuals [14]. Theoretical

* Corresponding author. Fax: +82 2 3410 2719.
E-mail address: kimjw@skku.edu (J.-W. Kim).

0888-7543/$ - see front matter © 2007 Published by Elsevier Inc.
doi:10.1016/j.ygeno.2007.11.001

research as well as empirical research has shown that a good ensemble can be more accurate than the best single predictor in the ensemble. Still, the ensemble approach does not always guarantee an improvement on the prediction. For a successful application of the ensemble approach, two conditions should be satisfied: first, the combined predictors should be more accurate than a random guessing for prediction and second, the errors made by the predictors should be mutually uncorrelated or negatively correlated [15]. To satisfy these criteria, we can divide data into two datasets and then calculate the prediction accuracy and the ratio error relationship of the predictors for the first set.

The first aim of this study is to evaluate schematically the performance of eight representative TSS/promoter predictors and to compare them on an identical computing platform and in an identical experimental environment. This will provide direction on how to choose a promoter/TSS predictor. The second aim is to apply the ensemble approach to the TSS prediction to improve the prediction accuracy and lower false prediction rates. We utilized three representative ensemble methods for that

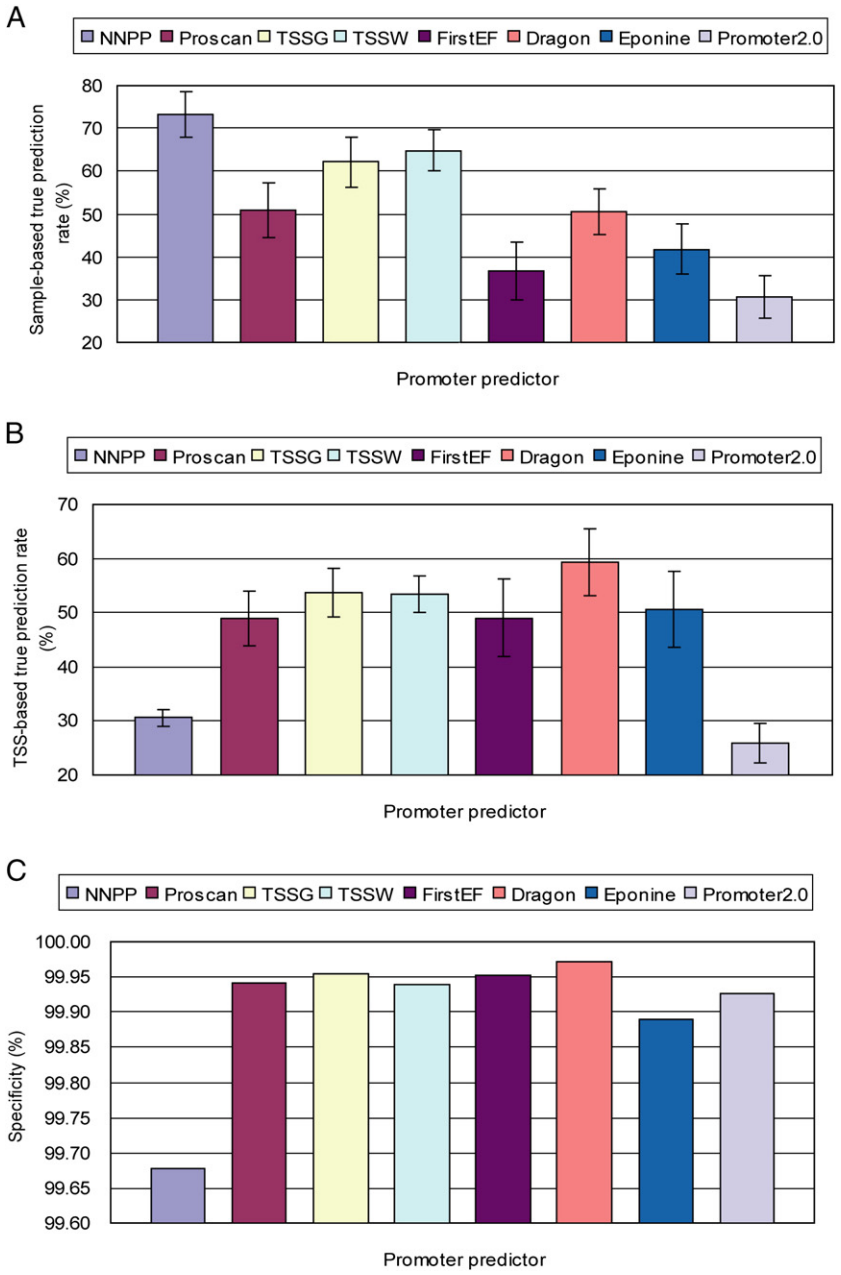

Fig. 1. The performance of individual promoter predictors for the training dataset. (A) Even though NNPP showed the best performance for the sample-based true prediction rate, (B) NNPP showed a relatively low TSS-based true prediction rate. This finding suggests that NNPP increases the sample-based true prediction rate by predicting many candidate TSSs per sequence. TSSG and TSSW showed high prediction performance for both the sample-based true prediction rate and the TSS-based true prediction rate. Dragon was superior for the TSS-based true prediction rate compared to the other predictors. (C) For specificity, all of the predictors showed high accuracy (>99.6%).

purpose—the majority voting approach, the weighted voting approach, and the Bayesian approach. Furthermore, we have modified the weighted voting and the Bayesian approach to increase the true prediction rate of promoter prediction. By defining a cutoff value and by including the predicted TSSs greater than the cutoff value, we were able to obtain more than one predicted TSS because of the ensemble prediction. We then examined the optimal cutoff value for the best prediction by changing the value of the cutoff point.

All of the ensemble approaches were superior in prediction accuracy to the eight individual predictors as seen in the experimental results. The false prediction rate of the ensemble predictor is lower than the false prediction rate of the single predictors. The weighted voting produced the best prediction performance among the proposed ensemble methods. The ensemble combining seven predictors was superior in the true prediction rate to the one combining three or five predictors. The results suggest that when more individual predictors are combined, the predictive ability of the ensemble predictor is improved.

Results and discussion

Evaluation of eight TSS/promoter predictors

We evaluated the accuracy of the eight TSS/promoter predictors to find the approximate location of TSS. If a predictor did not output an explicit TSS, we used the 3′ end of the predicted promoter region as the predicted TSS. We regarded a predicted TSS as correct if it was within the range from 200 bp upstream to 100 bp downstream of the experimentally defined TSS [12,16]. We evaluated the eight representative promoter predictors on preprocessed training datasets. To evaluate the general performance of the predictors, we repeated the 300 independent experiments. Since the promoter predictor outputted several predicted TSSs for each sample, we defined two prediction rates to measure the performance of the predictor. We defined the sample-based true prediction rate as the number of correctly predicted samples over the number of total samples, and we defined the TSS-based true prediction rate as the number of correctly predicted TSSs over the number of total predicted TSSs. The former rate is a measure of the probability that at least one of the possible promoters/TSSs predicted for each sequence by the method is within the true promoter region, and the latter rate is a measure of the probability that each predicted promoter/TSS is within the true promoter region. The sample-based true prediction rate is used for measuring the performance of the promoter prediction program, and the TSS-based true prediction rate is used for selecting the programs to combine. We calculated the average and the standard deviation of the true prediction rate and the specificity for the Eukaryotic Promoter Database (EPD) training dataset (see Methods) (Fig. 1). Specificity indicates the percentage of base pairs of the true negatives per the negative base pair regions. Comparing the performance of the predictors, we found that TSSG and TSSW showed the best performance for the sample-based true prediction rate and TSS-based true prediction rate. Although NNPP showed the highest sample-based true prediction rate of the eight predictors, it also showed a relatively low TSS-based true prediction rate. This finding indicates that the possibility of identifying the true promoter was inflated, as NNPP predicted many promoter predictions. Because Promoter2.0 did not surpass the accuracy of random-guessing prediction, we did not include it as a member of the ensemble predictors. Based on the TSS-based prediction rate of the single predictor, we chose combined predictors with high TSS-based true prediction rates.

Error correlation of a pair of predictors

As the error correlation of the predictors can affect the performance of the ensemble predictor, it is desirable, before applying the combination, to confirm that the predictors to be combined are less error-correlated. We calculated the error correlation in the case of the k-predictor combinations ($k=3$, 5, 7, and 8). Although the ratio of the error correlation of the eight-predictor combination was lower than the other combinations (Table 1), the experimental results showed that the seven-predictor combination predicted the TSSs more accurately than the other combinations. The reason for the seven-ensemble predictor being able to detect the TSSs with the highest prediction accuracy is that Promoter2.0, a member of the eight-ensemble predictor, inaccurately predicted the TSSs and affected the ensemble results. This finding suggests that the ensemble approach is useful if the combined predictors provide relatively accurate predictions. Therefore, even though the ratio of the error correlation of the eight-ensemble predictor was the lowest of all of the combinations, the prediction rate of the eight-ensemble predictor was lower than the prediction rate of the seven-ensemble predictor. Compared with the three-predictor combination, the five-predictor combination had a low ratio of error correlation and thus, the five-predictor ensemble had a relatively high predictive accuracy.

Optimal ensembles of multiple predictors

As the weighted voting and Bayesian approach output was only one predicted TSS because of the ensemble, their use showed a low false prediction rate along with a low true prediction rate. This result is related to the tradeoff between the false prediction rate and the true prediction rate for a possible cutoff value. An appropriate compromising point of the tradeoff

Table 1
Average ratio of the error relationship of the combined predictors

Combined predictors	φ_e^a	φ^{*b}	φ_e/φ^{*c}
Three predictors	0.255050	0.188947	1.349849
Five predictors	0.280868	0.222691	1.261245
Seven predictors	0.256171	0.207699	1.233376
Eight predictors	0.269979	0.233563	1.155915

[a] φ_e denotes the calculated error correlation of the combined predictors.
[b] φ^* denotes the expected error correlation when all the combined predictors made errors in a statistically independent manner.
[c] φ_e/φ^* denotes the ratio of the error relationship of the combined predictors.

should be chosen, as missing the optimal or true solution is as critical as taking suboptimal or false solutions. We examined how the prediction rate changed as the voting cutoff value varied for the 300 resampled datasets. For the seven-predictor-combined weighted voting method with a cutoff value α ranging from 0.5 to 1.0, the TSS-based true prediction rate ranged from 46.1 to 57.7% and the sample-based true prediction rate ranged from 67.6 to 87.0% for the training datasets. The inclination of the curve gets steepest around a cutoff of 0.7. The steep inclination means that one can obtain a higher sample-based true prediction rate with a small loss of the TSS-based true prediction rate.

Table 2 shows the comparison of the performance of the individual predictors and the ensemble predictors for the EPD test dataset. We defined the false positive rate (1 − specificity) as the number of base pairs of the false positives per the negative base pair regions. As a result of the ensemble of the individual predictors, we showed that the use of the ensemble approach was superior to the use of the best single predictor (TSSG/TSSW). The most powerful predictor in the sample-based prediction was the modified weighted voting ensemble with seven individual predictors. It produced a sample-based true prediction rate of 80.4%, while the best single predictors produced 65.1%. The use of the weighted voting ensemble with five single predictors (Proscan, TSSG, TSSW, Dragon, and Eponine) was superior for the TSS-based true prediction rate and false-positive rate compared to the use of the other predictors. The false positive rate of the use of the ensemble approach was lower than determined by the use of the single predictors. The individual predictors frequently missed true TSSs even though true TSSs existed in the sequences. The nondetection rates of the individual predictors were greater than 10%, except for NNPP. All the ensemble methods of the combined seven predictors had very low nondetection rates (Table 2). Comparing the ensemble methods, we observed that the application of the weighted voting was superior for the prediction performance to the application of the majority voting and Bayesian approach. The true prediction rate of the ensemble tended to increase as the number of combined predictors increased from three to seven, with the exception of from seven to eight. As the experimental results showed, we expected the prediction performance of the ensemble would converge at

Table 2
Comparison of the performance of the predictors and the ensemble predictors in the test dataset

Program	Sample-based true prediction (sensitivity, %)	TSS-based true prediction (%)	False positive rate (1 − specificity, bp)	Nondetection rate (%)
Single predictor				
NNPP	73.2±5.3	30.7±1.6	1/310	1.3±0.4
Proscan	49.5±5.2	49.5±5.2	1/1739	28.7±4.8
TSSG	62.4±5.9	54.1±4.7	1/2266	21.2±4.4
TSSW	65.1±4.9	53.7±3.4	1/1677	20.3±3.8
FirstEF	36.8±6.7	36.8±6.4	1/2130	36.8±6.4
Dragon	50.5±5.4	59.2±6.2	1/3451	15.2±3.1
Eponine	41.7±5.7	50.8±7.2	1/924	46.3±5.2
Promoter2.0	30.5±5.0	25.9±3.6	1/1376	13.3±3.0
Majority voting ensemble				
Three predictors	70.4±6.1	58.2±6.9	1/1993	7.3±2.8
Five predictors	71.3±6.9	61.6±5.8	1/2310	5.7±1.4
Seven predictors	76.6±4.7	52.9±5.0	1/1552	**0.0**±0.0
All predictors	75.1±5.0	53.3±4.3	1/1548	**0.0**±0.0
Weighted voting ensemble				
Three predictors	65.1±11.0	62.7±9.0	1/3063	7.3±2.8
Three predictors–mod.	69.3±6.5	59.2±6.5	1/2259	7.3±2.8
Five predictors	64.9±8.7	***66.9***±9.3	***1/3710***	5.7±1.4
Five predictors–mod.	72.2±8.0	60.9±5.0	1/2179	5.7±1.4
Seven predictors	67.9±6.8	58.0±8.5	1/2361	**0.0**±0.0
Seven predictors–mod.	***80.4***±5.3	52.1±4.2	1/1270	**0.0**±0.0
All predictors	66.9±6.0	58.4±6.9	1/2445	**0.0**±0.0
All predictors–mod.	79.0±4.8	52.6±4.4	1/1308	**0.0**±0.0
Bayesian ensemble				
Three predictors	51.7±9.4	60.4±11.9	1/3538	14.5±5.5
Three predictors–mod.	55.8±10.1	58.0±10.8	1/2729	14.5±5.5
Five predictors	54.1±10.0	60.9±12.7	1/3456	11.2±3.2
Five predictors–mod.	60.4±9.5	57.8±10.3	1/2374	11.2±3.2
Seven predictors	56.0±11.0	56.4±10.6	1/2771	0.8±0.5
Seven predictors–mod.	66.8±12.1	52.7±7.2	1/1680	0.8±0.5
All predictors	53.8±12.0	54.8±12.0	1/2598	0.8±0.5
All predictors–mod.	65.8±10.1	50.6±6.5	1/1569	**0.0**±0.0

We have defined the sensitivity as the percentage of samples with a correct promoter, the TSS-based true prediction as the percentage of the correctly predicted TSSs, the false positive rate as the number of base pairs of false positives per negative base pair regions, and the nondetection rate as the percentage of samples without a predicted promoter.

around seven predictors, even though more individual predictors were added to the ensemble predictor.

Comparison with other ensemble approaches

There have been other combining methods using an ensemble of several individual promoter predictors as reported in previous studies [12,13]. One method called CONPRO combined existing single methods including TSSG, TSSW, NNPP, Proscan, and PromFD for predicting promoters [12]. If three predictions fall in a 100-bp region, this is considered a consensus prediction in the method. The other method examined all possible combinations of predictions of the individual programs using two simple rules [13]. The algorithmic difference between the other methods and our method is that the same weight is given to all individual predictors in the other methods while a different weight is given to each individual predictor in our method.

The CONPRO correctly detected the promoters for about 71–73% of human genes with a known mRNA, promoters for about 63–65% of human genes with a known complete coding region, promoters for about 37–38% of human genes with known 3′ ESTs, and promoters for about 37–58% of human genes with known 5′ ESTs [12]. We have summarized the results of the previous studies in Table 3 and compared them with our results. The result of CONPRO shown in Table 3 is the performance of the predictors on the data of human genes with known 5′ ESTs [12]. The nondetection rate of our method was significantly reduced from 24.3 to 0.0%, while the nondetection rate of the CONPRO was slightly reduced from 45.1 to 38.2%. The use of EnsemPro resulted in an increase of 26.2% in the sample-based true detection rate and the use of CONPRO resulted in an increase of 18.6% in the sample-based true detection rate.

Table 3 also shows the best result of the combining method proposed by Bajic et al. [13]. These investigators examined all possible combinations of eight different prediction programs on chromosomes 4, 21, and 22. They defined the sample-based true prediction rate as the positive predictive value in their study. In the sample-based prediction, the sample represented a sequence of one gene in their definition. We compared the sample-based true prediction rate of our method with the positive predictive value of the best combination (i.e., a combination of Dragon, Eponine, FirstEF, and McPromoter) of all possible combinations in the previous study [13]. The respective gene is counted as a true positive when one or more predictions fall in the region [−2000, +2000] relative to the reference TSS location as described in the study [13], while the positive window region is defined as [−200, +100] in the present study and in a study by Liu and States [12]. The sample-based true prediction rate of the combining method in the Bajic et al. study [13] was increased 14.2% from 62.7 to 76.9%, compared with the average prediction rate of the use of the combined programs. The use of EnsemPro resulted in an increase of 26.2% in the sample-based true detection rate and the use of the method described in the Bajic et al. study [13] resulted in an increase of 18.6% of the sample-based true detection rate.

Conclusions

In this study, we proposed the use of an ensemble TSS/promoter predictor to improve prediction performance. The ensemble approach generally works well if the combined predictors are more accurate than random guessing and they are not error-correlated. We showed that our method can identify true TSSs with a low nondetection rate and high confidence and is highly accurate. The combined predictor has its own characteristics to explore a range of solutions with the use of its different algorithm.

Furthermore, we improved the prediction accuracy of the ensemble predictor by modifying its voting scheme. Adopting more predicted TSSs by adjusting the cutoff value increased the sample-based true prediction accuracy while it also decreased the TSS-based true prediction rate. When examining the sample-based true prediction rate and the TSS-based true prediction rate with changes of the cutoff value, we found that a cutoff value of 0.7 was appropriate for ensemble prediction

Table 3
Comparison of the prediction performance of the method in this study with those of previous studies

Program	Liu et al. [12]		Bajic et al. [13]	EnsemPro	
	Sample-based true prediction (sensitivity, %)	Nondetection rate (%)	Sample-based true prediction (%)	Sample-based true prediction (sensitivity, %)	Nondetection rate (%)
NNPP	40.2	30.4	Not combined	73.2±5.3	1.3±0.4
Proscan	24.5	73.5	Not combined	49.5±5.2	28.7±4.8
TSSG	51.0	36.3	Not combined	62.4±5.9	21.2±4.4
TSSW	34.3	38.2	Not combined	65.1±4.9	20.3±3.8
PromFD	46.1	47.1	Not combined	Not combined	Not combined
FirstEF	Not combined	Not combined	39.4	36.8±6.7	36.8±6.4
Dragon	Not combined	Not combined	64.8	50.5±5.4	15.2±3.1
Eponine	Not combined	Not combined	67.3	41.7±5.7	46.3±5.2
McPromoter	Not combined	Not combined	79.2	Not combined	Not combined
Average	39.2	45.1	62.7	54.2±5.6	24.3±4.0
Ensemble result	57.8	38.2	76.9	**80.4**±5.3	**0.0**±0.0

We compared the sample-based true prediction rate of the method in this study with the rates of the best combination of the methods in previous studies.

because at this value we could greatly increase the sample-based true prediction rate and insignificantly decrease the TSS-based true prediction rate for the cross-validation training datasets.

If more individual predictors were available, we might be able to improve the performance of the ensemble predictor, and the proposed ensemble promoter prediction method can be easily extended to combine more predictors. However, the results we generated showed that the effectiveness of increasing the number of the individual predictors gets weaker as the number of individual predictors increases. This is because the difference in the ratio of the error correlations decreases as the number of the combined predictors increases.

For future work, we will test the proposed method (EnsemPro) on the longer (>1.5 kb) sequence data by applying the sliding window method. As described in the study by Liu and States [12], about 80% of TSS fall within a region 10 kb upstream of the 5′ end for most exons in the gene. In terms of system performance evaluation, we will include data from the DataBase of Transcriptional Start Sites in future work, since more and more genes are known to have multiple TSSs [19]. Finally, we will also select newly developed promoter prediction programs as members of the ensemble predictor based not only on their performances, but also on their error correlation.

Methods

Fig. 2 describes the entire scheme of the evaluation process of the prediction performance of the proposed method. First, we downloaded 400 human genome sequences from the EPD. After preprocessing the data, we randomly partitioned it into two exclusive sets. We then evaluated eight individual predictors using the training dataset and selected the best *k*-predictors according to their rank based on the true prediction rate ($k=3$, 5, 7, and 8). We evaluated the combining results with the three ensemble approaches for the test dataset. We repeated this cycle 300 times to prevent any bias of the sampling dataset, and we estimated the generalized prediction performance.

EPD datasets

The EPD is an annotated nonredundant collection of experimentally characterized eukaryotic Pol II promoters [17]. The underlying promoter definition of the EPD was that of a TSS. Since the data in the EPD was collected from biologically confirmed results found in the scientific literature, we utilized them as the training and test datasets. First, we downloaded from the EPD ftp site 1871 human genome sequences that included TSSs and were 2.3 kb in length. We discarded any sequence containing missing base pairs from the raw data and chose the 400 DNA sequences without missing base pairs for the experimental dataset. We next randomly extracted subsequences of 1.5 kb in length to allow a true TSS to be located in a random point on the sequence, by referencing the length of the test sequences that were used in other studies [12,16]. We randomly divided the 400 preprocessed data into two exclusive datasets—one that was used for training a model and another that was used for testing the results. We finally generated 300 different training and test datasets so that each dataset had 200 nonredundant sequences for twofold cross-validation. We experimented with 300 independent tests using the training/test datasets for estimating the generalized performance, and we evaluated the TSS prediction performance.

Promoter prediction methods

There are scores of TSS predictors available on the World Wide Web (Internet). We chose eight of these predictors to combine their outputs for an ensemble as they have been actively maintained and been widely used. As they were developed using different algorithms and mathematical architecture, they have discriminative features, and it is these features that maximize the effectiveness of an ensemble. The individual methods used in this study are summarized as follows.

NNPP (Neural network promoter prediction)

NNPP is trained on TATA and Inr using the neural network algorithm, which allows variable lengths between them. It produces a predicted TSS [4].

Proscan (PromoterScan)

Proscan examines both the TATA box weight matrix and the density of transcription factor binding sites, and it compares them to the promoter recognition profile derived from a comparison of the promoter to nonpromoter

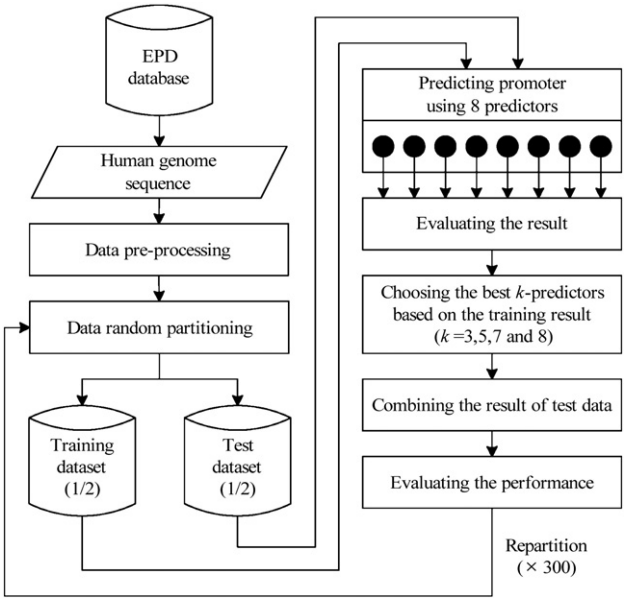

Fig. 2. Schematic illustration of the evaluation of the prediction performance (for details, see Methods).

Table 4
Relationship between correlation and statistical independence

$\varphi_e > \varphi^*$	Positively correlated	Dependent errors
$\varphi_e = \varphi^*$	Uncorrelated	Independent errors
$\varphi_e < \varphi^*$	Negatively correlated	Dependent errors

primate sequences. It produces either a TSS or a 250-bp window that represents a core promoter [5].

TSSG/TSSW

These two methods use the same underlying algorithm, which is the linear discriminant function based on the TATA box score, the triplet preferences around the TSS, the hexamer frequencies in consecutive upstream 100-bp regions, and transcription factor binding sites. It produces a list of transcriptional elements as promoter predictions [6].

FirstEF

FirstEF recognizes structural and compositional features such as CpG islands, promoter regions, and first splice-donor sites by using discriminant functions. It uses different models to predict CpG-related and non-CpG-related first exons [8].

Dragon promoter finder

This method first sorts the sequences according to CpG rich or CpG poor, then the sequences are passed through three sensors: promoter, exon, and intronic sequences. This method uses an artificial neural network to find the predicted promoter region [9].

Eponine

Eponine models use a collection of positioned constraints, and each one is represented by a DNA weight matrix. It combines the relevance vector machine with elements of a Monte Carlo sampling approach [10].

Promoter2.0

Promoter2.0 uses a neural network–genetic algorithm. Based on the conserved sequences and the conserved distances, it discriminates between promoter and nonpromoter sequences [11].

Ensemble methods

Majority voting ensemble

The majority voting ensemble is one of representative ensemble methods that can combine the outputs of multiple predictors. Majority voting has some advantages in that it does not require any previous knowledge or any additional complex computation for decisions. Where a target sequence is divided as exclusive m regions, $s_i(\text{predictor}_j)$ is 1 if the jth predictor predicts the i region as a TSS region; otherwise it is 0; and if c_e is an index of the predicted TSS region upon which the combined predictors agree, the majority voting is defined as follows:

$$c_e = \arg\max_{1\le i\le m}\left(\sum_{j=1}^{k} s_i(predictor_j)\right) \tag{1}$$

Weighted voting ensemble

A poor predictor can interrupt the enhancement of the performance of the majority voting ensemble because the majority voting ensemble gives the same weight to all predictors. Weighted voting reduces the effect of the poor predictor by giving a different weight to the predictor based on the performance of each predictor. We used the predictive rate of each predictor as the weight. Where w_j is the weight of the jth predictor, weighted voting is defined as follows:

$$c_e = \arg\max_{1\le i\le m}\left(\sum_{j=1}^{k} w_j s_i(predictor_j)\right) \tag{2}$$

As the weighted ensemble predictor produces only one predicted TSS of the maximum voting value, other possible true TSSs can be missing. One can select more than one predicted TSS by adding other predicted TSSs (c_m) greater than the predefined voting threshold of the modified weighted voting. Where α is the cutoff value and v_i is the weighted voting value of ith region, we have defined the voting threshold as the maximum voting value multiplied by α. In this study, we examined the sample-based true prediction rates and TSS-based true prediction rates by changing α from 0.5 to 1.0 for the training datasets and then finally determined the optimal value of α based upon the experimental results:

$$c_m = \left\{ c | v_c > \alpha \times \max_{1\le i\le m}(v_i),\ v_i = \sum_{j=1}^{k} w_j s_i(predictor_j) \right), 1\le c\le m\} \tag{3}$$

Bayesian voting ensemble

While the majority voting ensemble combines predictors without any preknowledge about them, the Bayesian ensemble uses the error possibility of each predictor; this method combines predictors with different weights by using the previous knowledge of each predictor. Where k-predictors are combined, $c(\text{predictor}_j)$ is the predicted TSS region of the jth predictor, and $p(c_i)$ is the probability that the ith region is a true TSS region, and the class of the Bayesian voting is defined as follows:

$$\begin{aligned} c_e &= \arg\max_{1\le i\le m}\left(\sum_{j=1}^{k} p(c_i | c(predictor_j))\right) \\ &= \arg\max_{1\le i\le m}\left(\sum_{j=1}^{k} \frac{p(c(predictor_j)\cap c_i)}{p(c(predictor_j))}\right) \end{aligned} \tag{4}$$

We assume that $p(c(\text{predictor}_j))$ is constant because the position of the true TSS is randomly decided when the dataset is generated. We add the predicted TSSs (c_m) as a result of the modified Bayesian ensemble predictor where α is the cutoff value and v_i is the voting value of ith region:

$$c_m = \left\{ c | v_c > \alpha \times \max_{1\le i\le m}(v_i),\ v_i = \sum_{j=1}^{k} \frac{p(c(predictor_j)\cap c_i)}{p(c(predictor_j))}, 1\le c\le m \right\} \tag{5}$$

Error correlation

We use the equation suggested by Ali [18] to evaluate the error correlation of the combined predictors. To apply the equation to our problem, we simplified the equation, as there is only one true TSS per single sequence. Where k-predictors are combined, $c(\text{predictor}_j)$ is the class of the jth predictor, and c_{true} is the true class of the sample, the degree of the error correlation ϕ_e is defined as follows:

$$\phi_e = \frac{1}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq 1}^{k} p\left(c(predictor_i)\neq c_{true}, c\left(predictor_j\right)\neq c_{true}\right) \tag{6}$$

A higher value of ϕ_e correlates with an increased number of errors made by members of the ensemble:

$$\begin{aligned} \phi^* = \frac{1}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq 1}^{k} & p(c(predictor_i)\neq c_{true}) \\ & \times p\left(c\left(predictor_j\right)\neq c_{true}\right) \end{aligned} \tag{7}$$

The error correlation ϕ_e would be ϕ^* if all the members of the ensemble produced errors in a statistically independent manner. If ϕ_e is less than or equal to ϕ^*, the members of the ensemble are negatively correlated or uncorrelated as shown in Table 4.

Table 5
Example of the relationship of the prediction results

Predictor 1	Predictor 2	φ_e	φ^*	φ_e/φ^*	Relationship
0 0 0 0 1 1 1 1	0 0 0 0 1 1 1 1	0.5	0.25	2	Positively correlated
0 0 0 0 1 1 1 1	1 1 0 0 1 1 0 0	0.25	0.25	1	Uncorrelated
0 0 0 0 1 1 1 1	1 1 1 1 0 0 0 0	0.0	0.25	0	Negatively correlated

Table 5 shows an example of the error relationship of the two predictors. In the example, 1 means the predictor correctly predicts the true class and 0 means the predictor falsely predicts the true class.

Acknowledgments

We thank M.G. Reese for providing NNPP, D.S. Prestridge for Proscan, V.V. Solovyev and A. Salamov for TSSG and TSSW, R.V. Davulury for FirstEF, V.B. Bajic for Dragon promoter finder, T.A. Down and T.J.P. Hubbard for Eponine, and S. Knudsen for Promoter2.0. This work was supported by a National Research Laboratory Grant from the Korea Institute of Science and Technology Evaluation and Planning, Republic of Korea.

References

[1] W.W. Wasserman, A. Sandelin, Applied bioinformatics for the identification of regulatory elements, Nat. Genet. 5 (2004) 276–287.
[2] U. Ohler, G. Liao, H. Niemann, G.M. Rubin, Computational analysis of core promoters in the Drosophila genome, Genome Biol. 3 (2002) (RESEARCH0087.1–0087.12).
[3] T. Werner, Models for prediction and recognition of eukaryotic promoters, Mamm. Genome 10 (1999) 168–175.
[4] M.G. Reese, Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome, Comput. Chem. 26 (2001) 51–56.
[5] D.S. Prestridge, Predicting Pol II promoter sequences using transcription factor binding sites, J. Mol. Biol. 249 (1995) 923–932.
[6] V.V. Solovyev, A. Salamov, The Gene-Finder computer tools for analysis of human and model organism genome sequences, in: T. Gaasterland, et al., (Eds.), Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, CA, 1997, pp. 294–302.
[7] V.V. Solovyev, I.A. Shahmuradov, PromH: promoters identification using orthologous genomic sequences, Nucleic Acids Res. 31 (2003) 3540–3545.
[8] R.V. Davulury, I. Grosse, M.Q. Zhang, Computational identification of promoters and first exons in the human genome, Nat. Genet. 29 (2001) 412–417.
[9] V.B. Bajic, S.H. Seah, A. Chong, G. Zhang, J.L.Y. Koh, V. Brusic, Dragon promoter finder: recognition of vertebrate RNA polymerase II promoters, Bioinformatics 18 (2002) 198–199.
[10] T.A. Down, T.J.P. Hubbard, Computational detection and location of transcription start sites in mammalian genomic DNA, Genome Res. 12 (2002) 458–461.
[11] S. Knudsen, Promoter2.0: for the recognition of Pol II promoter sequences, Bioinformatics 15 (1999) 356–361.
[12] R. Liu, D.J. States, Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling, Genome Res. 12 (2002) 462–469.
[13] V.B. Bajic, S.L. Tan, Y. Suzuki, S. Sugano, Promoter prediction analysis on the whole human genome, Nat. Biotechnol. 22 (2004) 1467–1473.
[14] N.C. Condorcet, Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix, Imprimerie Royale, Paris, 1785.
[15] T.G. Dietterich, Ensemble methods in machine learning: lecture notes in computer science, in: J. Kittler, F. Roli (Eds.), Proceedings of the First International Workshop on Multiple Classifier Systems, Springer-Verlag, London, 2000, pp. 1–15.
[16] J.W. Fickett, A.G. Hatzigeorgiou, Eukaryotic promoter recognition, Genome Res. 7 (1997) 861–878.
[17] R.C. Périer, T. Junier, P. Bucher, The eukaryotic promoter database EPD, Nucleic Acids Res. 26 (1998) 353–357.
[18] K. Ali, On the link between error correlation and error reduction in decision tree ensembles, Univ. of California Irvine Tech. Report, 1995.
[19] Y. Suzuki, R. Yamashita, S. Sugano, K. Nakai, DBTSS, DataBase of Transcriptional Start Sites: progress report 2004, Nucleic Acids Res. 32 (2004) D78–D81 (database issue).